(19) Europäisches Patentamt
European Patent Office
Office européen des brevets

(11) Publication number: **0 364 180 A2**

(12) EUROPEAN PATENT APPLICATION

(21) Application number: 89310277.2

(51) Int. Cl.5: G06F 15/401

(22) Date of filing: 06.10.89

(30) Priority: 11.10.88 US 255338

(43) Date of publication of application:
18.04.90 Bulletin 90/16

(84) Designated Contracting States:
BE DE FR GB IT LU NL

(71) Applicant: NeXT INC.
900 Chesapeake Drive
Redwood City California 94063(US)

(72) Inventor: Hawley, Michael J.
665 Woodland Avenue
Menlo Park California 94025(US)

(74) Representative: Hartley, David
Withers & Rogers 4 Dyer's Buildings Holborn
London, EC1N 2JT(GB)

(54) Method and apparatus for indexing files on a computer system.

(57) A method and apparatus for automatically indexing and retrieving files in large computer file systems is provided. Keywords are automatically extracted from files to be indexed and used as the entries in an index file. Each file having one of the index entries as a keyword is associated in the index with that keyword. If a file is to be retrieved, and its content but not it name or location is known, its keywords are entered and its identifying information will be displayed (along with that of other files having that keyword), facilitating its retrieval.

EP 0 364 180 A2

keyword, one can tell the order in which each file became associated with that keyword. The "dbAddIndex" routine then ends at step 25.

FIG. 3 is a flow diagram of the "findManFile(s)" routine of Appendix C which is exemplary of routines that can be used to retrieve files that have been indexed. At step 30, the index or indices to be searched are specified by the user or other program calling the routine. At step 31, the routine checks to see if all indices to be searched have been searched. If so, the routine ends at step 32. Otherwise, the next index is searched at step 33. If at step 34 the search text is found as a keyword in the index being searched, then the file or files with which it is associated are added to a reference list along with all identifying information and weights. The routine then returns to step 31 to see if any more indices need to be searched. Once the routine has ended at step 32, the reference list is presented to the user or other program that called the "findManFile" routine.

Hardware System

While the present invention may advantageously be implemented on nearly any conventional computer system, an exemplary hardware system 400 on which the present invention is implemented is shown in FIG. 4.

FIG. 4 shows a preferred embodiment of a hardware system 400 implementing the present invention as part of a computer system. In FIG. 4, system 400 includes CPU 401, main memory 402, video memory 403, a keyboard 404 for user input, printer 405, and mass storage 406 which may include both fixed and removable media using any one or more of magnetic, optical or magneto-optical storage technology or any other available mass storage technology and in which the files to be indexed and searched are stored (the files can be entered from keyboard 404 or directly into mass storage 406 on removable media; if system 400 is part of a network of computer systems, the file system might include all or part of the mass storage available on other systems on the network). These components are interconnected via conventional bidirectional system bus 407. Bus 407 contains 32 address lines for addressing any portion of memory 402 and 403. System bus 407 also includes a 32 bit data bus for transferring data between and among CPU 401, main memory 402, video memory 403, and mass storage 406. In the preferred embodiment of system 400, CPU 401 is a Motorola 68030 32-bit microprocessor, but any other suitable microprocessor or microcomputer may alternatively be used. Detailed information about the 68030 microprocessor, in particular concerning its instruction set, bus structure, and control lines, is available from MC68030 User's Manual, published by Motorola Inc., of Phoenix, Arizona.

Main memory 402 of system 400 comprises eight megabytes of conventional dynamic random access memory, although more or less memory may suitably be used. Video memory 403 comprises 256K bytes of conventional dual-ported video random access memory. Again, depending on the resolution desired, more or less such memory may be used. Connected to a port of video memory 403 is video multiplex and shifter circuitry 408, to which in turn is connected video amplifier 409. Video amplifier 409 drives cathode-ray tube (CRT) raster monitor 410. Video multiplex and shifter circuitry 408 and video amplifier 409, which are conventional, convert pixel data stored in video memory 403 to raster signals suitable for use by monitor 410. Monitor 410 is of a type suitable for displaying graphic images having a resolution of 1120 pixels wide by 832 pixels high.

The reference lists produced by the invention can be stored in mass storage 406, displayed to the user on monitor 410, or printed out on printer 405.

Thus it is seen that a method and apparatus for automatic indexing of files in a computer system are provided. One skilled in the art will appreciate that the present invention can be practiced by other than the described embodiments, which are presented for purposes of illustration and not of limitation, and the present invention is limited only by the claims which follow.

Claims

1. For use in a computer system having a mass storage file system, apparatus for automatic indexing of files stored in said file system, said apparatus comprising:
means for maintaining at least one index file on said file system, said at least one index file containing a date of most recent indexing and, for each of at least some of said stored files, information regarding a date of last updating for that file, and an association of that file with at least one keyword; and
means for automatically updating said at least one index file when a new file is added to said file system and when an existing file is updated, said automatic updating means comprising:
means for, for each file, comparing said date of most recent indexing and said date of last updating for that file,
means for determining if said date of last updating of that file is later than said date of most recent

indexing, and
means for updating said keyword association information upon a determination by said determining means that said date of last updating of that file is later than said date of most recent indexing.

2. The apparatus of claim 1 wherein:
each of said stored files has associated therewith a file type; and
said automatic updating means further comprises:
means for, for each file, examining said file type, and
means for, based on said file type, extracting from said file keywords and information concerning the relative occurrences of said keywords.

3. The apparatus of claim 1 further comprising means for retrieving said files based on said keyword association information.

4. For use in a computer system having a mass storage file system, a method for automatic indexing of files stored in said file system, said method comprising:
maintaining at least one index file on said file system, said at least one index file containing a date of most recent indexing and, for each of at least some of said stored files, information regarding a date of last updating for that file, and an association of that file with at least one keyword; and
automatically updating said at least one index file when a new file is added to said file system and when an existing file is updated, said automatic updating step comprising:
for each file, comparing said date of most recent indexing and said date of last updating for that file,
determining if said date of last updating of that file is later than said date of most recent indexing, and
updating said keyword association information upon a determination that said date of last updating of that file is later than said date of most recent indexing.

5. The method of claim 4 wherein:
each of said stored files has associated therewith a file type; and
said automatic updating step further comprises:
for each file, examining said file type, and
based on said file type, extracting from said file keywords and information concerning the relative occurrences of said keywords.

6. The method of claim 4 further comprising retrieving said files based on said keyword association information.

80/7/222 (Item 222 from file: 350)
Derwent WPIX

0005128930 *Drawing available*
WPI Acc no: 1990-117688/
XRPX Acc No: N1990-091207
Computer system file indexing apparatus - has automatic keyword extraction and file indexing, storing and retrieval based on keywords
Patent Assignee: NEXT COMPUTER INC (NEXT-N); NEXT INC (NEXT-N)
Inventor: HAWLEY M J

Patent Family (2 patents, 8 countries)

Patent Number	Kind	Date	Application Number	Kind	Date	Update	Type
EP 364180	A	19900418	EP 1989310277	A	19891006	199016	B
CA 1318404	C	19930525	CA 615087	A	19890929	199326	E

Priority Applications (no., kind, date): US 1988255338 A 19881011

Patent Details

Patent Number	Kind	Lan	Pgs	Draw	Filing Notes	
EP 364180	A	EN				
Regional Designated States,Original	BE DE FR GB IT LU NL					
CA 1318404	C	EN				

Alerting Abstract EP A
The appts. **automatically** indexes files and provides for their **retrieval** based on **keywords**. The **keywords** used for filing are **automatically** extracted from a file on the basis of their frequency of use, or weighting, as compared their frequency of use in a **relevant** language domain. The choice of reference domain can be defined by the user or selected on the basis of the file **type**. A filter mechanism is provided to remove **words** which are not suitable for use as **keywords**, such as pronouns in natural language and reserved **words** in a **computer language**.
The files are indexed either when they are updated or created, or alternatively may be updated on a background basis. Access to the files for either user **queries** or **application** program access cna be based on **relevant keywords**.
USE/ADVANTAGE - Allows files in large filing system to be selected on basis of **keywords** generated **automatically**. @(10pp Dwg.No.4/4)@

Title Terms /Index Terms/Additional Words: COMPUTER; SYSTEM; FILE; INDEX; APPARATUS; **AUTOMATIC**; **KEYWORD**; EXTRACT; STORAGE; **RETRIEVAL**; BASED

Class Codes

International Patent Classification

IPC	Class Level	Scope	Position	Status	Version Date
G06F-015/401			Main		"Version 7"
G06F-015/40			Secondary		"Version 7"

File Segment: EPI;
DWPI Class: T01
Manual Codes (EPI/S-X): **T01-J05B**

Original Publication Data by Authority

Canada

Publication No. CA 1318404 C (Update 199326 E)
Publication Date: 19930525
Assignee: NEXT COMPUTER INC (NEXT-N)
Inventor: HAWLEY M J
Language: EN
Application: CA 615087 A 19890929 (Local application)
Priority: US 1988255338 A 19881011
Original IPC: G06F-15/401(A)
Current IPC: G06F-15/401(A)

EPO

Publication No. EP 364180 A (Update 199016 B)
Publication Date: 19900418
Verfahren und Vorrichtung zum Indizieren von Dateien in einem Computer
Method and apparatus for indexing files on a computer system
Methode et dispositif pour indexer des fichiers dans un systeme ordinateur
Assignee: NeXT INC., 900 Chesapeake Drive, Redwood City California 94063, US (NEXT-N)
Inventor: Hawley, Michael J., 665 Woodland Avenue, Menlo Park California 94025, US
Agent: Hartley, David, Withers & Rogers 4 Dyer's Buildings Holborn, London, EC1N 2JT, GB
Language: EN
Application: EP 1989310277 A 19891006 (Local application)
Priority: US 1988255338 A 19881011
Designated States: (Regional Original) BE DE FR GB IT LU NL
Original IPC: G06F-15/40
Current IPC: G06F-15/40
Original Abstract: A method and apparatus for automatically indexing and retrieving files in large computer file systems is provided. Keywords are automatically extracted from files to be indexed and

used as the entries in an index file. Each file having one of the index entries as a keyword is associated in the index with that keyword. If a file is to be retrieved, and its content but not it name or location is known, its keywords are entered and its identifying information will be displayed (along with that of other files having that keyword), facilitating its retrieval.

Claim: The appts. automatically indexes files and provides for their retrieval based on keywords. The keywords used for filing are automatically extracted from a file on the basis of their frequency of use, or weighting, as compared their frequency of use in a relevant language domain. The choice of reference domain can be defined by the user or selected on the basis of the file type. A filter mechanism is provided to remove words which are not suitable for use as keywords, such as pronouns in natural language and reserved words in a computer language.

The files are indexed either when they are updated or created, or alternatively may be updated on a background basis. Access to the files for either user queries or application program access cna be based on relevant keywords.